Theories of Democratic Backsliding

Edoardo Grillo[1], Zhaotian Luo[2], Monika Nalepa[3], and Carlo Prato[4]

[1]Department of Economics and Management, University of Padova, Italy

[2,3]Department of Political Science,The University of Chicago, Chicago, IL, USA

[4]Department of Political Science, Columbia University, New York, NY, USA

October 22, 2023

Abstract

We review recent contributions to the modeling of democratic backsliding. We organize existing theories according to (1) the source of constraints on the executive (vertical or horizontal restrainers), and (2) the target of backsliding (electoral manipulation or aggrandizement), and then use them as scaffolding for a meta-model of democratic backsliding. This meta-model allows us to describe and compare the premises and insights of this scholarship. We further apply our 2-dimensional classification to over thirty empirical papers and we show how theoretical assumptions can guide research design by clearly highlighting the scope conditions of different theories of backsliding. We conclude the review by highlighting open issues for future research.

Acknowledgments: The authors are grateful to Hanna Folsz, Marko Klašnja, Daniel Markovits, Anne Meng, Chit Basu and Adam Przeworski for their excellent comments, as well as Daniel Markovits and Simon Siskel for outstanding research assistance. All remaining mistakes are the authors' responsibility.

1 Theorizing backsliding

Throughout the 20th century, democratization was the modal form of regime transition. Unsurprisingly, this direction of regime change was also the focus of scholars and policymakers (Geddes, 1999; Huntington, 1993; O'donnell et al., 2013; Przeworski, 1991, 2000; Kuran, 1991). Scholarship on democratic breakdown concentrated on the dynamics of coups d'etát (De Bruin, 2018; Svolik, 2012), with less attention devoted to other forms of breakdown (though see Linz and Stepan, 1978). In the 21st century, however, democratic backsliding has become increasingly common, prompting a shift in scholars' attention (Gandhi, 2019; Waldner and Lust, 2018; Gamboa, 2017; Bermeo, 2016; Haggard and Kaufman, 2021).

Whether democratization is now less frequent than democratic backsliding is an open question (Little and Meng, 2023), but the sheer volume of academic output devoted to democratic backsliding suggests that scholars are still struggling to converge on a way of understanding this phenomenon. A recent wave of formal theoretical work, however, highlights that the gradual dismantling of democracy is more than just the flip side of democratic consolidation.

The goal of this review is to take stock of this body of work and improve its accessibility to researchers, particularly those with more empirical agendas. Specifically, the review will clarify how formal models explain democratic backsliding and how they answer questions such as "Does political polarization contribute to backsliding?", "Are more popular and electorally secure incumbents more likely to attempt to weaken democracy?" and "When do institutional checks and balances provide an effective bulwark against democratic dismantling?", among others.

A comparative analysis of this body of work shows that the answers to the above questions depend on two fundamental features of the polity under consideration: (1) the source of constraints on the executive, and (2) the main consequence of the executive's attempts to weaken democratic institutions. We begin our discussion by presenting a

meta-model predicated on these two features. We believe that this exercise facilitates the discussion of each specific contribution. By allowing the reader to map these theories of backsliding into the more general meta-model, it increases the overall legibility of these models and helps contextualize their implications. In the absence of the organizing meta-model, some of these empirical implications may appear ambiguous or even contradictory.

In the remainder of this review, we apply the same classification scheme that informed our meta-model to organize more than thirty empirical articles. This exercise underscores how the assumptions of theoretical papers can be translated into scope conditions that empiricists should consider when thinking about research designs and generating expectations.

We conclude by raising some questions that we believe the next wave of scholarship on democratic backsliding should address.

2 Defining our Terms

Scholars have often used the term "democratic backsliding" interchangeably with "populism", "autocratization," and "democratic erosion." Yet these terms capture distinct concepts. The difference between backsliding and populism is largely one of domains. Whereas one would use the term "populist" to describe the attributes of leaders and policy stances, it hardly applies to institutions or constitutions (which are systems of interrelated institutions). Populist attempts to undermine established elites (e.g., judges) or political organizations (e.g., parties) often result in the "disfigurement of representative democracy" (Urbinati, 2019), but very often anti-elitism is a goal in itself (Vachudova, 2021) whereas the undermining of democracy is merely collateral damage.

Contrary to "democratic backsliding," "autocratization" is most commonly used in the context of regimes that are already autocratic, to denote a contraction in the size of the coalition supporting the regime, that is, a process of de-institutionalization. While it is

tempting to think of more institutionalized regimes as more democratic, Meng (2020)'s work documents how institutionalized autocracies are more durable, and hence less likely to democratize, than personalistic autocracies. Hence, we believe that autocratization and democratic backsliding are processes that take place on separate tracks.

Our avoidance of the term "democratic erosion" largely stems from the fact that it suggests a spontaneous process that is not driven by the choices of goal-oriented actors, a premise that all theories discussed here share.

In this piece, "democratic backsliding" denotes the process of removing constraints on accountability for democratically elected executive leaders (Waldner and Lust, 2018; Przeworski, 2019; Levitsky and Ziblatt, 2018). Democratic backsliding weakens institutions and rules that determine who can vote, who can run for office, who can compel the executive to provide information, and who can block unilateral executive action. We can classify these constraints based on whether

(a) they prevent the executive from manipulating elections, or

(b) they restrict the executive's discretion in various policy areas.

We can also classify these constraints based on whether

(a) they originate with the electorate or other large groups of citizens (vertical constraints);

(b) they originate with other elite actors: another government branch, party leaders, or media outlets (horizontal constraints).

Based on this two-dimensional way of thinking about democratic backsliding, we organize the formal models we review (as well as the empirical articles presented in Section 6) into a two-by-two matrix, Table 1.

The dichotomy between electoral manipulation versus aggrandizement is distinct from the more familiar one between office and policy motivation often featured in theories of

	Constraints	
Type	vertical	horizontal
el. manipulation	Svolik, 2020 Gratton and Lee, 2023 Luo and Przeworski, 2023	Helmke et al., 2022 Miller, 2021 Hollyer et al., 2023
aggrandizement	Grillo and Prato, 2023 Howell and Wolton, 2018 Chiopris et al., 2021	Howell et al., 2023

Table 1: A 2-dimensional taxonomy of recent formal theories of democratic backsliding

electoral institutions. Our dichotomy is not based on differences in political actors' *motives* but on *tools* available to them and/or *outcomes* achievable with unilateral action. Even incumbents who are entirely motivated by policy outcomes might resort to electoral manipulation to achieve their goals. Similarly, aggrandizement can be a tool to improve one's electoral security (e.g., if illiberal measures are sufficiently popular within a primary electorate).

3 The meta-model

A polity is composed of three agents: an Incumbent ("he," the chief executive), an institutional Restrainer ("she," e.g. the opposition, an elite political actor, or the judiciary), and the Electorate ("it," the citizens). All agents have preferences over a space $\mathbb{R}$. The space $\mathbb{R}$ may capture ideological positions, but also degrees of authoritarianism.

The interaction between the three actors takes place in two phases: a *democratic subversion phase* and a *policy-making phase*. Initially, the Incumbent holds office and a status quo policy normalized to 0 is in place. In the democratic subversion phase, the Incumbent can initiate a reform that can weaken democracy (the reform can extend his tenure, change electoral laws to reduce the odds of his defeat, or curtail media freedom). In response, the Restrainer decides whether to oppose this initiative, which can either succeed or fail. Next, in the policy-making phase, the Incumbent proposes a policy from a set that depends on the success of the initiative in the subversion phase. Subsequently, the Electorate decides

whether to continue supporting the Incumbent. The Incumbent is reelected with a probability that depends on the Electorate's support. If reelected, the Incumbent implements his policy. If the Incumbent is not reelected, the status quo policy prevails.

Below, we describe the democratic subversion and policy-making phases in more detail. Table 2 summarizes the variables, parameters, and functions of the model.

3.1 Democratic subversion phase

In this phase, the Incumbent chooses the intensity of subversion, captured by an initiative $\rho \in \mathbb{R}_+$, where $\rho = 0$ denotes no subversion. If subversion does not occur or is not successful, no backsliding occurs. If successful, subversion results in two possible forms of backsliding. First, it results in the loosening of the constraints the Incumbent faces when choosing the policy. Absent successful subversion, in the policy-making phase the Incumbent can choose a policy from a set $\mathcal{Z}_0 = [-z, z]$. Successful subversion can expand this set to $\mathcal{Z}_\rho = [-(z+\rho), z+\rho]$. Note that z captures the initial institutional constraints that limit the executive's ability to change the status quo. ρ, meanwhile, measures the additional discretion that subversion grants. Second, successful subversion makes the Incumbent's position in office more stable. Without support from the Electorate, the Incumbent maintains office with a probability $\varphi(\rho) \in [0, 1]$, where φ is an increasing function.

These two effects of successful subversion can operate simultaneously or in isolation. If successful subversion has no effect on the Incumbent's policy authority, we set $\mathcal{Z}_\rho \equiv [-z, z]$. If successful subversion has no effect on the probability of retaining power, we set $\varphi(\rho) \equiv \phi$, where $\phi \in [0, 1]$ is a constant.

The Restrainer can oppose the Incumbent's initiative ($b = 1$) or not ($b = 0$). The action $b = 1$ captures the use of legislative (e.g., passing bills) and non-legislative (e.g., organizing public demonstrations) actions by the opposition to reverse the consequences of the initiative, or a court striking the initiative down. If the Restrainer opposes the Incumbent's initiative, she pays a cost $c > 0$, but the initiative fails with probability $q \in$

$(0, 1)$. The parameter q thus captures the *strength of the Restrainer*. With complementary probability, $1-q$, the initiative passes despite the Restrainer's opposition, and subversion is successful. If the Restrainer does not oppose, the initiative passes, subversion is successful, and the Restrainer incurs no cost.

We denote by $r \in \{0, \rho\}$ the outcome of the subversion phase: $r = \rho$ if subversion is successful (the Restrainer does not oppose or she opposes but the Incumbent's initiative passes); $r = 0$ if subversion is unsuccessful (the Restrainer opposes and the Incumbent's initiative fails) or there is no subversion.

3.2 Policy-making phase

In the policy-making phase, the Incumbent proposes a policy change from the status quo (normalized to 0) to a policy $\xi \in \mathcal{Z}_r$, where r is the outcome of the subversion phase.

The Incumbent then runs for reelection and the Electorate decides whether to support him. If the Electorate supports him, the Incumbent retains office. If the Electorate does not support him, the Incumbent stays in power with probability $\varphi(r)$ and loses office with probability $1 - \varphi(r)$. If the Incumbent maintains office, the policy platform ξ is implemented. If the Incumbent loses office, platform ξ is not implemented and the status quo remains in place. The term $x \in \{0, \xi\}$ denotes the implemented policy: $x = \xi$ if the Incumbent maintains office and $x = 0$ if the Incumbent loses office.

The Electorate's decision to support the Incumbent can thus depend both on his policy initiatives, as in traditional models of electoral accountability and electoral control, and on whether and how the Incumbent's previous actions *change* the nature of accountability by either strengthening his hold on office and/or expanding his policy authority.

3.3 Preferences

Each agent i has policy preferences represented by a quadratic loss relative to her *preferred policy* $\beta_i \in \mathbb{R}$, where $i = I(\text{ncumbent}), R(\text{estrainer}), E(\text{lectorate})$.[1] When policy x is implemented, agent $i \in \{I, R, E\}$ gets a policy payoff $u_i(x) = -(x - \beta_i)^2$.

In addition to his policy payoff, the Incumbent gets a payoff equal to $\Pi \geq 0$ from retaining office, and the Restrainer pays a cost equal to $c \geq 0$ for opposing democratic subversion. The actual *objective function* of each agent—reflecting the goals they pursue—can take a specific form in the models we review. Where appropriate, we describe these specifics in each model's dedicated section.

Choice variables	
$\rho \in \mathbb{R}_+$	Democratic subversion
$\xi \in \mathcal{Z}_\rho$	Policy choice
$b \in \{0, 1\}$	Restrainer opposition
Parameters	
z	Institutional constraints
q	Strength of the Restrainer
c	Cost of opposition
Π	Incumbent's value from office
β_i	Agent i's bias, $i \in \{I, R, E\}$
Functions and Outcomes	
$r \in \{0, \rho\}$	Executive aggrandizement
$x \in \{0, \xi\}$	Implemented policy
$\mathcal{Z}_r = [-(z + r), z + r]$	Incumbent's policy authority
$\varphi(r) \in [0, 1]$ (increasing)	Electoral manipulation

Table 2: Summary of choice variables, parameters, and functions of the meta-model

4 A taxonomy of recent models

An advantage of formal models—as with any exercise in abstraction—is their flexibility in covering many seemingly different situations. In this section, we take advantage of

[1] While a fully congruent Restrainer corresponds to the case $\beta_R = \beta_E$, a Restrainer's preferred policy may differ from the Electorate's.

this flexibility and nest existing models of democratic backsliding in the framework we introduced in Section 3.

A general disclaimer applies to our analysis. Some of the narratives of the papers we discuss do not fit neatly within the boundaries of our two-by-two matrix. In these instances, we used the formal model as our guide. For instance, the criterion we use to classify a paper as aggrandizement is not whether attempts to subvert democracy enable more extreme policies, but whether they expand the Incumbent's ability to change policy outcomes.

4.1 Vertical restrainers and electoral manipulation

The first set of models studies settings where the Incumbent manipulates elections and voters are the main defense against these attempts. This corresponds to the northwest quadrant of Table 1.

In Svolik (2020), an Incumbent runs for reelection against a non-strategic Challenger. The Electorate is split into two groups: informed (share $1 - alpha$) and uninformed (share α). Informed voters value democracy and have preferences over implemented policy. Uninformed voters do not have preferences over policy and can be swayed to support the Incumbent through manipulative actions. The Incumbent chooses the extent $\rho \in (0, 1)$ of this manipulation, which has two effects. First, it affects the reelection probability of the Incumbent. Second, it modifies the composition of the Incumbent's support. Because informed voters share a common value for fair elections (referred to as "civic virtue"), higher manipulation increases Incumbent support among uninformed voters and decreases it among informed ones.

Informed voters differ in their policy preferences. A typical informed voter j has a utility equal to

$$u_{E,j}(x) = -(x - \beta_{E,j}) - \rho^2,$$

where the term $-\rho^2$ reflects his civic virtue, namely the dislike for electoral manipulation. The informed voters' ideal points $\beta_{E,j}$ are distributed in the interval $[-z, z]$ according to a cumulative density function with point masses at the extremes (recall that in our meta-model, the parameter z represents the institutional constraints *before* the Incumbent has a chance to subvert democracy). The Incumbent and the Challenger differ in their ideal policies: The Incumbent is biased to the right ($\beta_I > 0$) and the Challenger is biased to the left ($\beta_R < 0$). The policy space does not depend on the degree of manipulation: $\mathcal{Z}_\rho \equiv [-z, z]$.[2]

In Svolik (2020), voters with extreme ideal policies tolerate electoral manipulation to obtain a policy closer to their ideal point. When the share of these extreme voters is sufficiently large, the Incumbent's optimal strategy is to propose a policy biased to the right and choose a positive level of manipulation, which enhances his probability of retaining office.

Therefore, a more polarized Electorate (that is, more voters with extreme ideal points) weakens democratic institutions (leads to higher ρ, that is, weaker electoral accountability) and also leads to more extremism.

Observation 1. *The support for incumbents who engage in electoral manipulation increases with mass polarization.*

Voters do not need to value democracy *intrinsically* to be willing to defend it. As suggested by familiar intuitions from models of electoral accountability, voters may value democracy *instrumentally*, because it is easier to discipline and select incumbents that may be replaced when falling short of their promises. In this vein, Luo and Przeworski (2023) present a model where voters have no exogenous preference for democracy. Yet, they have preferences over the quality of (that is, the policy congruence with) their repre-

[2]Note that expressing Svolik's model in our framework allows us to see that this is a model of electoral manipulation and not aggrandizement because the set $\mathcal{Z}_\rho$ remains constant. At first sight, however, and based on the empirical section of the paper, one may reach a different conclusion.

sentatives, and they value democracy to the extent that it allows them to achieve better policy outcomes.[3]

In line with this idea, Luo and Przeworski (2023) derive this instrumental (common) *value of democracy* from a trade-off between improving current policy outcomes and retaining the ability to select future incumbents. Consider an incumbent with autocratic ambitions facing a challenger of lower quality: replacing the incumbent leads to less desirable policy outcomes today, but prevents electoral manipulation in subsequent periods by making it easier to replace today's challenger with better opponents in the future.

In their setting, the Incumbent can gradually tilt the playing field in his favor. This accumulated advantage increases the probability that the Incumbent retains office against the Electorate's will. In the notation of our meta-model, the Incumbent at time t can choose whether to *subvert* democracy ($\rho_t = 1$ versus $\rho_t = 0$). Subversion increases the Incumbent's probability of retaining office when the Electorate withdraws its support: $\varphi_t = (1 - \rho_t)\varphi_{t-1} + \rho_t\psi$, where ψ is a random variable distributed according to some arbitrary distribution with support inside $[\varphi_{t-1}, 1]$. In the first period, the Incumbent has no electoral advantage, $\varphi_0 = 0$.

Agents' ideal points capture preferences over a spatial policy dimension. The Electorate's ideal point is fixed at $\beta_E = 0$. The Incumbent's ideal point is to the right of the Electorate's: $\beta_I \in (0, 1)$. The identity and preferences of the Challenger, instead of being fixed, evolve stochastically. In each period t, the preferences of the Challenger align with those of the Electorate (that is, $\beta_{R,t} = 0$) with probability $\gamma \in (0, 1)$. With complementary probability, the Challenger has an ideal policy equal to $\beta_{R,t} = -1$, which is more extreme (in the opposite direction) than the Incumbent's.

When the Incumbent chooses to subvert (that is, when he accumulates electoral advantage, $\rho_t = 1$), and the Challenger is an extremist ($\beta_{R,t} = -1$), the Electorate faces

[3]We note that in place of "preferences over the policy", one could have "preferences over ideologies" or "preference over valence".

a policy-democracy trade-off. If it keeps supporting the Incumbent, its policy payoff in the current period is higher than if it withdrew the support. The Incumbent, however, accumulates additional electoral advantage and the Electorate's ability to replace him in the future with a Challenger that shares its ideal policy decreases.

In equilibrium, two backsliding pathways are possible. When the Incumbent is relatively popular (that is, β_I is sufficiently close to 0 or γ is sufficiently close to 0), the Incumbent always subverts democracy and the Electorate keeps supporting him. Because the Incumbent is likely to be better than the Challenger, the Electorate does not value the opportunity to replace him in the future. The Incumbent thus is emboldened to increasingly tilt the scales in his favor over time.

Conversely, when the Incumbent is relatively unpopular (that is, β_I is sufficiently large or γ is sufficiently close to one), he is so unlikely to retain the Electorate's support that manipulation becomes his only hope for retaining office—even as he anticipates voters to sanction him. Since voters expect manipulation in subsequent periods, they withdraw their support even when the Incumbent ends up facing less popular (more extreme) challengers. This, in turn, provides the Incumbent with further incentives to subvert democracy.

Along both backsliding pathways, the Incumbent can accumulate enough electoral advantage to almost fully insulate himself from the electoral process. In Luo and Przeworski (2023), democracy is sustainable—that is, voters sanction attempts to subvert democracy—only if the Incumbent is sufficiently unpopular *and* the Challenger is unlikely to be a better alternative.

Observation 2. *Very high and very low levels of incumbent electoral security are associated with backsliding.*

While Svolik (2020) shows that mass polarization, that is, deep policy disagreements among voters, can exacerbate the risk of democratic backsliding, its occurrence can also be prompted by "common value issues," such as national security. Gratton and Lee (2023)

highlight this possibility through an infinite horizon model of electoral accountability.

Gratton and Lee (2023) assume that the Electorate values liberty and democracy, but also confronts the risk of negative shocks (e.g., terrorist attacks, foreign invasions, or natural disasters). Shocks occur with some exogenous probability and *only authoritarian leaders can handle them*. The Electorate receives a noisy signal concerning the occurrence of these shocks (e.g. from the media), after which it must elect a liberal Incumbent or an authoritarian one. The elected Incumbent then subverts democracy ($\rho = 1$) if and only if he is authoritarian.

Democratic subversion expands the Incumbent's policy authority: $\mathcal{Z}_0 = \{0\} \subseteq \mathcal{Z}_1 = \{0, 1\}$.[4] The Electorate's ideal policy is $\beta_E = 0$ if there is no shock and $\beta_E = 1$ if there is a shock.

In this setting, the Electorate is willing to trade off liberty for security. It thus elects an authoritarian Incumbent if and only if the probability that the situation warrants authoritarian intervention is high enough. This happens when the expected loss from implementing policy $\xi = 0$ when the shock occurs is greater than the expected loss from implementing policy $\xi = 1$ when there is no shock.

Democratic subversion, however, also increases the probability that the Incumbent retains power by selectively censoring the information available to the Electorate. By manipulating and censoring information, the Incumbent can persuade the Electorate that the negative shock is sufficiently likely to justify his retention.[5] Electing an authoritarian Incumbent has thus both short- and long-term costs. In the short term, the Electorate incurs a loss when the authoritarian Incumbent implements policy $\xi = 1$ but there is no shock. In the long term, the Electorate suffers the consequences of weakened electoral accountability due to signal manipulation.

Gratton and Lee (2023) show how the probability of a negative shock, the Electorate's

[4] In our meta-model, this corresponds to normalizing $z = 0$, to economize on notation.

[5] Formally, the Incumbent can modify the information structure available to the Electorate.

value for democracy, and the informativeness of the signal available to the Electorate jointly determine the evolution of political regimes over time. If the probability of a shock is low enough, the Electorate appoints and retains a liberal Incumbent. In this case, liberal regimes are both stable and efficient. At the other extreme, if the probability of a shock is high, the Electorate appoints and retains an authoritarian Incumbent who can handle the impending crisis. In this case, authoritarian regimes are both stable and efficient: the authoritarian Incumbent does not need to censor information to convince the Electorate to retain him. Finally, when the probability of a shock is intermediate, the authoritarian Incumbent is eventually elected and he engages in censorship to retain power. In this scenario, although the Electorate would like to get rid of authoritarian Incumbents more often, censorship enables these regimes to last inefficiently long (possibly forever).

Observation 3. *Trading-off liberty for security can initiate backsliding. Illiberal incumbents selectively censor information to stoke the fear of security shocks, reinforcing the backsliding process.*

4.2 Vertical restrainers and policy aggrandizement

While maintaining the focus on the incumbent-restraining power of the electorate, in this section we discuss theories that apply to the many circumstances where successful democratic subversion expands the Incumbent's power rather than (or in addition to) securing his hold on power.

Chiopris et al. (2021) posit an Electorate facing an Incumbent who wants to expand his policy authority. The Electorate, however, is uncertain regarding the true intentions of the Incumbent. In particular, the Incumbent is one of two types: a Closet Autocrat with probability $\alpha \in (0, 1)$ or an Ideologue with probability $1 - \alpha$. The type of the Incumbent is his private information.

The Incumbent initially proposes an institutional reform $\rho \in [0, 1]$ and then decides

which policy ξ to propose. A higher ρ increases the Incumbent's power through one of two channels: first, by expanding policy authority from $\mathcal{Z}_0 = \{0\}$ to $\mathcal{Z}_\rho = [0, \rho]$; and second, a higher ρ enables the Incumbent to pursue further aggrandizing goals. Both the Closet Autocrat and the Ideologue prefer the most extreme policy ($\beta_I = 1$ regardless of the Incumbent's type), but they differ in their desire of aggrandizing. The Ideologue only harbors extreme policy preferences and he has no intention to exploit institutional reforms beyond ordinary policy changes. The Closet Autocrat, instead, is a true subverter: he regards institutional reforms as instrumental to his overarching goal of power aggrandizing (that is, also values ρ for its own sake).

The Electorate dislikes aggrandizing: if the Closet Autocrat is in power, the Electorate suffers a loss equal to $-\rho^2$, while it suffers no such cost if the Ideologue is in office. On top of this preference against aggrandizing, the Electorate has spatial preferences with an ideal policy equal to $\beta_E \in (0, 1)$. The overall utility of the Electorate is thus $u(x, \rho) = -(x - \beta_E)^2 - \rho^2$. After observing ρ, the Electorate decides whether to keep supporting the Incumbent or to withdraw its support. If the Electorate withdraws its support, the Incumbent loses office and a (nonstrategic) Challenger with ideal policy $\beta_R = 0$ replaces him. Since the Challenger is not a strategic player in the game, his ideal point can also be interpreted as the status quo or reversion point.

Chiopris et al. (2021) show that the Electorate supports the Incumbent over the Challenger if the institutional reform ρ is lower or equal than a threshold $\overline{\rho}(\alpha, \beta_E)$ that is decreasing in the probability that the Incumbent is a Closet Autocrat (α) and in the ideal policy of the Electorate (β_E). The properties of the threshold $\overline{\rho}(\alpha, \beta_E)$ have an intuitive interpretation: the Electorate tolerates more extensive institutional reforms when Closet Autocrats are less likely, and when the Electorate itself prefers significant changes in the policy domain.

A particularly interesting type of equilibrium is the one where both types of Incumbent pool and propose the most extensive reform that is compatible with the Incumbent's

reelection, $\rho = \overline{\rho}(\alpha, \beta_E)$. In this equilibrium, if the Electorate has moderate policy preferences (β_E is low), its belief that the Incumbent is a Closet Autocrat needs to be low for the Incumbent to be reelected. However, as its policy preferences become more extreme, the Electorate reelects the Incumbent even when there is a non-negligible chance that he is a Closet Autocrat. In this latter scenario, the Electorate may experience regret upon learning that a Closet Autocrat has gained power, in line with the evidence in Svolik (2021).

We observe two differences between Chiopris et al. (2021)'s contribution and Svolik (2020)'s. First, democratic backsliding is possible even when the Electorate exhibits very strong commitments to democracy. Second, when Closet Autocrats are sufficiently common, the Electorate may accept suboptimal policies to prevent aggrandizement from happening. In practical terms, the Electorate may only reelect and Incumbent whose institutional reforms are so moderate that they prevent the Electorate's ideal point from being adopted. This loss in the policy dimension is the toll the Electorate pays to prevent severe backsliding.

Observation 4. *Voters may enforce limits to institutional reform that separate Closet Autocrats from Ideologues. These limits, however, prevent the Incumbent from adopting voters' preferred policy when elected.*

The notion that voters trade away ideological proximity in order to prevent democratic backsliding is the focus of the empirical section of Chiopris et al. (2021).

Informational asymmetries can also interact with voter's emotions, with subtle consequences. Grillo and Prato (2023) explicitly consider the role of fear. In their model, the Incumbent decides between subverting democracy ($\rho = 1$) or not ($\rho = 0$). The baseline model is one of pure aggrandizement: subversion only expands the set of policies available in the policy-making phase (that is, $\mathcal{Z}_1 = [-1, 1] \supseteq \mathcal{Z}_0 = \{0\}$, while the function φ is constant). The key innovation is that beyond an ideal policy equal to $\beta_E = 0$, the

Electorate has *reference-dependent* utility: a policy x generates a psychological gain when $u_E(x)$ is above the Electorate's *reference point*, which is its expected policy payoff after the Incumbent's initial choice of subversion but before his final policy choice, $\mathbb{E}[u_E(x)]$. Similarly, policy x generates a psychological loss when $u_E(x)$ is below $\mathbb{E}[u_E(x)]$. In words, the Electorate's expectations about the Incumbent's future behavior shape its emotional reaction to his *actual* behavior. The support the Incumbent receives from the Electorate is thus increasing in the psychological gain associated with $u_E(x)$ exceeding $\mathbb{E}[u_E(x)]$ and decreasing in the psychological loss associated with $u_E(x)$ falling short of $\mathbb{E}[u_E(x)]$.

The Incumbent's aggrandizement is interpreted here as enabling changes away from the status quo 0 and towards positive values of the policy x. Since $\beta_E = 0$ and $\beta_I > 0$, these changes benefit the Incumbent and hurt the Electorate. More specifically, the Incumbent's ideal policy can be moderately high ($\beta_I \in (0, 1]$) or extreme ($\beta_I > 1$). The ideal policy of the Incumbent is his private information and is determined by some exogenous probability. An Incumbent with an extreme ideal policy chooses democratic subversion ($\rho = 1$) and the most extreme policy ($\xi = 1$), regardless of the Electorate's reaction.

When the Electorate sees that the Incumbent initiates democratic subversion (that is, he chooses $\rho = 1$), its expectations regarding policy become pessimistic, fearing that policy $\xi = 1$ will get implemented. Any less extreme choice of the Incumbent in the policy-making phase ($\xi < 1$) comes thus as a relief and, due to reference dependence, can increase the Incumbent's electoral support. When reference dependence is sufficiently strong, the Incumbent with moderate ideal policy exploits this *fear-and-relief* mechanism. He subverts democracy only to back down and use the Electorate's sense of relief to attract its support.

The opportunity to exploit this fear-and-relief mechanism originates in the emotional response of the Electorate to the initiatives of the Incumbent that could weaken democracy. Grillo and Prato (2023) show that this has two notable implications. First, rather than preventing democratic subversion, stronger pro-democracy values in the electorate might *encourage* it. Second, a more polarized Electorate might actually *discourage* incumbents

from engaging in democratic subversion: as voters' policy concerns are on average stronger than their democratic concerns, the fear-and-relief strategy is of limited electoral use in a highly polarized electorate.

Observation 5. *When voters exhibit reference dependence, subversion followed by policy moderation might enhance the electoral prospects of the executive relative to no subversion; mass polarization might decrease the likelihood of subversion.*

Although formal initiatives (e.g., institutional reforms) can set the ground for future authoritarian moves, they are not always a necessary step. Incumbents can also expand their power informally, through channels that are partially hidden from the general public. Howell and Wolton (2018) address this possibility. Their model features an Incumbent, who can propose a change in the status quo policy ($x = 0$) in one of two possible directions ($\xi \in \{-1, 1\}$). The policy change can occur through two different channels: *de jure* aggrandizement or *de facto* aggrandizement.

De jure aggrandizement takes place through institutional reforms ($\rho = 1$), is costly, and cannot be opposed by the institutional Restrainer ($\rho = r$). Furthermore, it permanently expands the Incumbent's policy authority ($\mathcal{Z}_1 = \{-1, 0, 1\} \supseteq \mathcal{Z}_0 = \{0\}$). *De facto* aggrandizement, in contrast, need not be preceded by institutional reforms, is costless, and the Restrainer can block it with an exogenous probability $q \in (0, 1)$. Finally, successful *de facto* aggrandizement allows the Incumbent to choose a specific policy, say $x = 1$, but it does not expand his policy authority in both directions: if the Incumbent also wants to implement policy $x = -1$, he needs to expand his authority in this other direction in a subsequent period.

Once the Incumbent chooses the type of aggrandizement, he runs against a Challenger.

Players' ideal policies are set equal to $\beta_I = 1$ for the Incumbent, $\beta_R = -1$ for the Challenger and $\beta_E \in (-1, 1)$ for the Electorate. The Incumbent's probability of reelection is increasing in β_E. When the preferences of the Electorate and of the Incumbent are far apart

(that is, when β_E is low enough), attempts to expand the Incumbent's authority come at a significant electoral cost. The Incumbent anticipates his replacement with the Challenger should he attempt to expand his authority and he thus refrains from aggrandizement.

When the ideal policies of the Electorate and of the Incumbent are closer to one another (that is, when β_E is sufficiently high), the Incumbent will seek power aggrandizement. The channel employed—de jure or de facto—depends on the Electorate's policy responsiveness. If the Electorate is highly responsive to policy, *de jure* aggrandizement improves the Incumbent's electoral odds. If the Incumbent loses the election, *de jure* aggrandizement would allow the Restrainer to implement policy $x = -1$. Given the high β_E, the Electorate wants to avoid this and it thus throws its support behind the Incumbent. But when the Electorate puts less weight on policy, the above reasoning collapses and the Incumbent chooses the cheaper *de facto* aggrandizement.

Observation 6. *De facto aggrandizement is more likely when voters put less weight on policy at the ballot. De jure aggrandizement is more likely when policy changes are salient to voters and become the key drivers of electoral results.*

4.3 Horizontal restrainers and electoral manipulation

Models that analyze the interaction between incumbents and vertical restrainers highlight that the Electorate is not always able to rein in authoritarian behaviors. While electorates in practice often lack the information, willingness, and ability to effectively perform their restraining role, there are other elite actors who, by virtue of being relatively less vulnerable to these shortcomings, might be able to do so.

In this vein, Miller (2021) considers a model where the Incumbent chooses between engaging in democratic subversion to preserve power ($\rho = 1$) or maintaining the status quo ($\rho = 0$). The Electorate cannot observe ρ. However, an institutional Restrainer, after observing an imperfect signal of ρ, can call on the Electorate to mobilize against the

Incumbent. The Electorate can either mobilize ($b = 1$) or not ($b = 1$).

If democratic subversion succeeds, the Incumbent secures office ($\varphi(1) = 1$) against the wishes of the Electorate. If the Electorate mobilizes and overthrows the Incumbent, the Restrainer replaces the incumbent ($\varphi(1) = 0$) and enjoys the value from office Π.[6] In all remaining cases, the incumbent retains power with probability $\varphi(1) = \gamma \in (0, 1))$ and loses it with probability $1 - \gamma$. The reform has no impact on the policy space ($\mathcal{Z}_1 = \mathcal{Z}_0$).

The Electorate mobilizes more easily if (i) it believes that the Restrainer calls for mobilization only when her signal tells her that the Incumbent chose $\rho = 1$, (ii) it believes that the Restrainer's signal about ρ is accurate, and (iii) it is not ideologically aligned with the Incumbent. Condition (i), however, should not be taken for granted: the Restrainer can "cry wolf" to replace the Incumbent in power and enjoy the value of office Π.

Miller (2021) shows that democracy is stable (that is, the Incumbent does not try to grab power) if and only if elections are sufficiently competitive—that is, when γ takes intermediate values and both Incumbent and Restrainer have a sufficiently high chance of winning the election in the absence of subversion. When the Incumbent's electoral chances are very thin (low γ), subversion is her only way to hold on to office. When the Incumbent's electoral chances are very high (high γ), the Restrainer will call for mobilization too often and, in equilibrium, the Incumbent will respond with frequent subversion.

This logic leads to a somewhat paradoxical result: increasing the Electorate's readiness to mobilize in the name of its democratic values can weaken the stability of democratic institutions. Facing a more responsive Electorate, the Restrainer is tempted to "cry wolf" often enough for the Incumbent to attempt power grabs.

Observation 7. *Power grabs are less likely when electoral contests are competitive and the electorate's readiness to mobilize is not too large.*

The opposition can also discipline the ruling party through dynamic incentives that

[6] While Miller (2021) considers an infinite horizon model in which Incumbent and Restrainer can alternate in power, we can capture its key intuitions with a simpler setting.

guarantee the self-sustainability of democratic institutions. This is the topic of Helmke et al. (2022) that study how parties can use, for instance, gerrymandering to subvert democracy.

In their model, two parties (A and B) alternate in holding executive office with equal probability in each period. When party A holds office, he can subvert democracy (which the authors interpret as "tilting the election") and increase his probability of retaining office from $1/2$ to $\varphi_A \in \left(\frac{1}{2}, 1\right)$. The value of φ_A represents party A's *ability to subvert democracy*. Similarly, when party B holds office, she can increase her probability of remaining in office in the next period from $1/2$ to $1 - \varphi_B$, where $\varphi_B \in \left(\frac{1}{2}, 1\right)$ represents party B's ability to subvert democracy.

In this repeated interaction, Helmke et al. (2022) provide conditions under which a *forbearance equilibrium* is possible. In this equilibrium, neither party subverts democracy and an even and fair alternation of power is sustainable. The forbearance equilibrium is supported by grim-trigger strategies: at the start of the game neither party subverts democracy while in office and they keep refraining from subversion as long as no subversion occurs. As soon as someone subverts, both parties start subverting democracy and keep doing so forever. The forbearance equilibrium places two conditions on parties' abilities to subvert democracy: first, these abilities must be sufficiently low (that is, both φ_A and φ_B are sufficiently close to $1/2$); second, parties need to be able to subvert democracy to a similar extent (that is, φ_A and $1 - \varphi_B$ are sufficiently close). The first condition guarantees that the temptation to subvert is not too high for either party. The second condition ensures that the threat of punishment for subversion is high enough to deter both parties.

Observation 8. *Symmetry in the parties' ability to subvert democracy prevents democratic backsliding through a mutual deterrence mechanism.*

While Miller (2021) and Helmke et al. (2022) focus on the restraining role of the opposition, scholars have also documented how the Incumbent's *own party* can perform this role

(Levitsky and Ziblatt, 2018). This idea is formally explored by Hollyer et al. (2023). The trade-off faced by the Incumbent's co-partisans is similar to the policy-democracy trade-off described in Svolik (2020): punishing an Incumbent who engages in electoral manipulation is valuable because it helps preserve democracy, but it hampers the pursuit of the party's programmatic goals by enhancing the opposition's electoral chances. The key argument of Hollyer et al. (2023) is that the cost of removing a charismatic Incumbent is high because he is very likely to win against the opposition. A charismatic Incumbent is thus less likely to be internally restrained by his own party. This, in turn, implies that charisma and backsliding should be positively associated, in contrast to both Luo and Przeworski (2023) and Miller (2021).

4.4 Horizontal restrainers and policy aggrandizement

In Miller (2021), it is the competition between the Restrainer and the Incumbent that limits the Restrainer's credibility and, consequently, her ability to prevent authoritarian reforms. In contrast, for Howell et al. (2023), the Restrainer is unable to restrain because she knows she might share the Incumbent's preferences in the future.

In Howell et al. (2023)'s model, the initial policy authority is equal to $\mathcal{Z}_0 = \{0\}$. The Incumbent can expand this policy authority through an institutional reform $\rho \in (0, 1)$: $\mathcal{Z}_\rho = [-\rho, \rho]$. The Restrainer can oppose the reform ($b = 1$) or refrain from doing so ($b = 0$). The Electorate does not play a relevant role and we can thus ignore it for the purposes of this discussion.

The Incumbent's ideal policy, β_I is high enough for him to always prefer choosing the most extreme policy available on the right. If the Restrainer does not oppose the reform, the Incumbent will implement $x = \rho$. When the Restrainer opposes the reform, instead, the policy reverts to the status quo, $x = 0$, and the Incumbent loses the opportunity to propose further institutional reform. The Restrainer's ideal policy, β_R, depends on a policy state that is realized after the Restrainer's choice of action. The state, similarly to Gratton

and Lee (2023), is interpreted as national a security threat, natural disaster, or health emergency. In times of emergency, the Restrainer agrees with the Incumbent's preferred course of action, while in normal times the Incumbent and the Restrainer disagree.

In light of the uncertainty surrounding β_R, opposing the institutional reform is risky for the Restrainer. When the Restrainer and the Incumbent have little or no conflicting interests (β_R is likely to be high), the Restrainer prefers the Incumbent choose a positive policy $x > 0$ rather than the status quo. Anticipating this, the Incumbent can propose an institutional reform $\rho > 0$ that the Restrainer will accept.

The main contribution of Howell et al. (2023) is to show that this mechanism unravels dynamically: once the Incumbent has expanded his policy authority to $[-\rho, \rho]$, he can expand it further to $[-\rho', \rho']$ with $\rho' > \rho$. The key feature behind this kind of authoritarian escalation is that the policy the Incumbent implements when the Restrainer opposes is the upper bound of the set feasible policies from the previous period.

Observation 9. *If executive aggrandizement can serve horizontal restrainers' future interests, they will not prevent democratic backsliding in the present.*

4.5 Additional considerations

Before discussing the empirical implications of these theories, we must mention that many separation-of-origins systems are explicitly designed around the interaction between vertical and horizontal restrainers. In presidential systems under divided government, the president is accountable to citizens both directly as well as indirectly, via congressional leaders with varying incentives to perform their restraining role. In semi-presidential systems, a directly elected president relies on a prime minister and a cabinet that needs the support of a legislative majority originating in separate elections. To date, the closest formal model capturing this dual restrainer situation is Miller (2021).

5 Empirical implications of the meta-model

We now take advantage of the unified analytic framework in Section 3 to derive and discuss a few key comparative implications about the determinants of the *risk of backsliding*, a term that captures both the probability and the expected severity of backsliding.

First, consider the effect of polarization in the Electorate, or *mass* polarization. In the language of our meta-model, this is captured by the difference between β_E of different voters, or the distance between β_E on the one side and β_R or β_I on the other side. The effect of mass polarization depends on the kind of backsliding enacted by the Incumbent. In the case of electoral manipulation (as in Svolik, 2020,) higher mass polarization increases the risk of backsliding. In the case aggrandizement, however, electoral polarization may (as in Chiopris et al., 2021) or may not (as in Grillo and Prato, 2023) increase the chances of backsliding.

The effect of *elite* polarization (the programmatic distance between the Incumbent's preferred policy and the Restrainer's one, captured by the distance between β_I and β_R) is also nuanced. It can decrease the risk of backsliding via aggrandizement (restrainers are more likely to oppose aggrandizement in Howell et al., 2023), but not under electoral manipulation (parties are less likely to sanction their own charismatic leaders in Hollyer et al., 2011).

A third implication concerns the Electorate's normative attachments to democratic values. Contrary to what conventional wisdom might suggest, strengthening voters' democratic values can *increase* the risk of backsliding, as shown, via different mechanisms, by both Grillo and Prato (2023) and Miller (2021). Frequently correlated with the strength of democratic values is citizens' belief that democratic backsliding is something that might not happen in their own polity. For instance, established democracies will have both high democratic values and be unlikely to suspect they are in danger of backsliding. Chiopris et al. (2021)'s model of aggrandizement with vertical restrainers implies that such polities

are more vulnerable to backsliding, other features held constant.

Fourth, the general contestability of elections acts as a stabilizing force: more competitive elections reduce the risk of backsliding in both Helmke et al. (2022) and Miller (2021). Competitive elections accomplish this by increasing the incentives of the Restrainer to oppose backsliding.

Finally, the effect of incumbents' personal popularity depends on the key set of actors constraining the Incumbent's behavior: under horizontal accountability, more popular incumbents are more likely to engage in backsliding (Hollyer et al., 2023), but this is not always the case under vertical accountability (Luo and Przeworski, 2023).

6 The Empirical Literature

In this section, we concentrate on recent empirical articles on democratic backsliding (papers with both formal and empirical components were discussed in Section 4). [7] Our goal is twofold. On the one hand, we hope that this piece can show empirical researchers how to select formal papers for their theoretical sections where they generate hypotheses. On the other hand, we hope to encourage formal modelers to pay closer attention to the interests of empirical researchers, by focusing, for instance, on the analysis of the implications that lend themselves to empirical explorations.

Despite the relative paucity of cross-citations, there is an affinity between the modeling assumptions of the theoretical papers and the scope conditions in the empirical literature. A paper that runs a conjoint experiment with potential voters should not draw as motivation for hypotheses on theories with horizontal restrainers. Likewise, an experiment with a vignette describing redistricting efforts or other forms of electoral manipulation should not be seeking justification for hypotheses in papers that focus on aggrandizement.

[7] There is a large literature in American Politics and American Political Development about support for civil liberties, which in some cases overlaps with the focus of this review (Green et al., 2011, e.g.,). Motivated by space constraints, our criterion for inclusion is a narrower and explicit focus on democratic backsliding, though we acknowledge the relevance of this literature for current scholarship on backsliding.

Our taxonomy, based on the target of backsliding (whether the incumbent is manipulating the electoral process in his favor or expanding his powers) and the key incumbent-restraining actor (voters or other elites) can be used to also organize empirical papers. The specific terms used to distinguish vertical from horizontal restrainers vary across articles. Some authors (e.g., Wunsch et al., 2023) refer to "the people" and "the elites," but it is important that research designs focusing on a certain type of restrainer preventing a certain type of backsliding draw on the appropriate theories.

Using the same classification as Table 1, Table 3 summarizes the empirical literature on democratic backsliding. Each paper highlighted in bold cites theories with assumptions (the kind of restrainer and the type of backsliding) that match the scope conditions of the empirical paper's research design. The papers that cite some formal theoretical articles, though not exactly those that match the scope conditions, are in regular font.

Out of 37 papers described in Table 3, ten focus on horizontal restrainers and, among them, only three consider aggrandizement alone (three more deal with both aggrandizement and electoral manipulation). All but two of these ten papers (Druckman et al., 2023 and Das, 2023, although the latter is primarily devoted to documenting electoral manipulation) use observational data. This is not surprising given the difficulties of performing experiments on elite actors. We note that the empirical articles that focus on horizontal restrainers either do not cite the theoretical literature, or cite papers with vertical rather than horizontal restrainers.

Out of the 27 papers with vertical restrainers, 11 focus on aggrandizement, and the remaining 15 focus on electoral manipulation alone (8), or both (8). Most of these papers (all but 4) use experimental data.

	Constraints	
Type	vertical	horizontal
el. manipulation	Van Noort (2022); Braley et al. (2023) Arceneaux and Truex (2022); Wuttke et al. (2023) Clayton et al. (2021);**Frederiksen (2022b)** Aarslew (2023); Arbatli and Rosenberg (2021)	Ziblatt (2017); Grumbach (2022) Druckman et al. (2023) *Das (2023)*
aggrandizement	**Svolik (2021)**;Kingzette et al. (2021) Bischof et al. (2023); Wuttke and Foos (2022) **Mazepus and Toshkov (2021)**; *Wuttke et al. (2023)* Rovny (2023);Wunsch et al. (2023);*Braman (2021)* **Grossman et al. (2022); Gidengil et al. (2022)**	Pérez-Liñán et al. (2019) Thompson (2021) *Benasaglio Berlucchi and Kellam (2023)*
both	Gandhi and Ong (2019); Krishnarajan (2023) **Simonovits et al. (2022); Orhan (2022)** **Svolik et al. (2023); Cinar and Nalepa (2022)** **Frederiksen (2022a); Saikkonen and Christensen (2023)**	Gibler and Randazzo (2011) Arriola et al. (2021) Pospieszna and Vetulanicegiel (2021)

Table 3: Empirical papers on democratic backsliding. The papers in italics cite none of the formal models we discuss in Section 4. Those in normal font cite some formal models from Table 1, but there is a mismatch between the location of these citations in the two-by-two matrix and the scope conditions of the empirical paper. In bold font, we highlight the papers that cite formal models matching their scope conditions.

With the exception of Gibler and Randazzo (2011), all the empirical articles in Table 3 were published after 2019, and thus do not clearly predate the models discussed in Sections 4 and 5. Although three papers (in italics) cite no paper from Section 4, a majority (in regular font) cite some of them, and as many as 11 cite the theories whose assumptions matched the empirical scope conditions. Out of these, we highlight three to illustrate how, ideally, one would incorporate theory into an empirical testing exercise.

6.1 Vertical restrainers, electoral manipulation (Svolik 2021)

The dominant form of empirical research on democratic backsliding is conjoint experiments. Respondents are confronted with either a pair of candidates with different ideological attributes but similar democratic attachments (control), or with a pair of candidates of whom one is associated with democratic subversion—typically, aggrandizement.

Svolik (2021) constitutes an exception. The paper exploits a natural experiment: the overturning of the 2019 Istanbul Mayoral elections by the dominant AKP party, which under the pretext of irregularities ordered a rerun. By comparing the behavior of voters in the first and second election—which features the same pair of candidates and the same electorate—one can obtain an empirical estimate of the electoral punishment associated with manipulation.

Svolik (2021) distinguishes between three possible channels of electoral punishment: switching (which involves voters who previously cast ballots for AKP switching to CHP), backlash (which involves differential increases in turnout to the benefit of CHP), and disenchantment (which involves differential decreases in turnout in favor of CHP). The evidence suggests that the latter two channels are relatively more important than the first.

This empirical exercise builds on the author's prior theoretical work, but it also mentions other papers from the west cells of Table 1, such as Chiopris et al. (2021); Grillo and Prato (2023) and Luo and Przeworski (2023). Yet, among those, only Luo and Przeworski (2023) directly theorizes the effects of manipulation.

The work that perhaps would have been most suitable for Svolik (2021)'s hypothesis, however, is the one by Gratton and Lee (2023), who explicitly model how voters interpret government's information deferentially depending on the context. Since the official reason for invalidating the election pointed to irregularities and "organized crime," it is entirely plausible that the electorate's response was driven by increased skepticism toward the government's claims about what Gratton and Lee (2023) call a negative shock.

6.2 Vertical restrainers, aggrandizement (Grossman et. al. 2022)

A good example of using theories of aggrandizement to explain how vertical restraining works is Grossman et al. (2022). In the US context, the paper employs survey experiments to elicit support for a "power grab," in which a governor inappropriately appoints a judge to a state supreme court. The authors' key interest is in the role of democratic values as a driver of opposition to democratic subversion. They draw on the theories from the south-west corner of Table 1—more specifically Grillo and Prato (2023) and Chiopris et al. (2021). Both choices are warranted by the focus on vertical restrainers and on the incumbent's aggrandizement goal (by means of undermining judicial independence). In addition, both of these works explicitly model democratic attachments.

Grossman et al. (2022) conceptualize these attachments (or the lack thereof) by contrasting three groups of voters: Majoritarians, Autocrats, and Militants. The former might fail to sanction power grabs not because they attach a negative value to democracy (as Autocrats do) or because they are willing to trade-off democratic erosion for policy accomplishments (as Militants do), but because they think of the incumbent as the embodiment of the will of the majority.

This majoritarian deference proves to be stronger than partisan attachments: voters fail to sanction even incumbents *from the opposing party*—as long as they have attracted majority support. We note that the very concept of majoritarian democratic values presupposes that fair elections are taking place. As such, this paper could hardly be rooted

in models of electoral manipulation.

6.3 Vertical restrainers, mixed motives (Frederiksen 2022)

Frederiksen (2022b) examines voters' willingness to sanction incumbents' electoral manipulation (e.g., via vote buying) in the UK when a third party (e.g., Liberal Democrats) becomes available. This study finds that the presence of a third party reshapes the electoral punishment against the incumbent party, as some voters no longer defect to the main opposition party but choose Liberal Democrats instead, without increasing the overall magnitude of the defection.

The authors' candidate choice conjoint experiment involves two levels of randomization: the presence of a third party and, conditional on the former, whether or not this party is a viable option (captured by past membership in the executive). When the incumbent attempts to manipulate elections, the presence of the third party reduces defections to the main opposition party by pro-democratic voters.

The empirical hypotheses for this paper draw on two articles in the north-west cell of Table 1: Svolik (2020) and Luo and Przeworski (2023). Invoking the former is justified because voters' readiness to punish subversion is affected by their spatial preferences and these, in turn, are affected by the emergence of a second opposition party. The somewhat surprising finding that the third party does not encourage more defections is actually in line with the intuitions developed by Luo and Przeworski (2023). Introducing a third party increases voter's perceived probability of a high-quality challenger. Holding the behavior of the incumbent fixed, this implies that the dynamic loss associated with electoral manipulation can be *lower* for a voter, since she believes she will be more likely to get rid of the incumbent in the future.

7 Conclusion

This article provides a structured overview of a recent formal theoretical literature on democratic backsliding. Its key contribution is to provide a meta-model, which allows one to organize recent contributions around two dimensions: the identity of the restrainer and the main target of backsliding. This exercise identifies the key premises that existing theories of democratic backsliding rely on. It makes these theories more accessible to empirical researchers, allowing them easily assess if the scope conditions implied by these theories are compatible with their empirical settings.

We hope that this review contributes to an empirical agenda of theoretically informed work on the occurrence of backsliding, with particular focus on the role of (i) ideological and affective polarization, (ii) the distribution of citizens' attitudes towards democracy, (iii) transparency, (iv) the relative popularity of the incumbent vis-á-vis the opposition, and (v) institutional checks and balances.

Our piece has limitations. As with all classification exercises, our taxonomy necessarily forces us to straight-jacket some research—especially the empirical articles—into rigid categories, with some loss of nuance. We take some comfort in the notion that one glaring gap in empirical research is hardly attributable to this rigidity. Table 3 suggests that there is not enough work on horizontal restrainers. This is troubling for two reasons. First, non-elite actors—voters and citizens—have considerably less frequent opportunities and fewer tools to exercise their restraining roles relative to elite actors. Second, focusing on vertical restrainers carries the risk of conflating majoritarianism (and its most radical interpretations) with democratic rule.

Going forward, methodological challenges notwithstanding, the integration between theories and empirical designs on horizontal restrainers is crucial, given that the accumulation of knowledge will necessarily have to proceed via the analysis of observational data.

References

Aarslew, L. F. (2023). Why Don't Partisans Sanction Electoral Malpractice? *British Journal of Political Science 53*(2), 407–423.

Arbatli, E. and D. Rosenberg (2021). United we stand, divided we rule: how political polarization erodes democracy. *Democratization 28*(2), 285–307.

Arceneaux, K. and R. Truex (2022). Donald Trump and the Lie. *Perspectives on Politics*, 1–17.

Arriola, L. R., J. DeVaro, and A. Meng (2021). American Political Science Review. pp. 82.

Benasaglio Berlucchi, A. and M. Kellam (2023). Who's to blame for democratic backsliding: populists, presidents or dominant executives? *Democratization*, 1–21.

Bermeo, N. (2016). On democratic backsliding. *Journal of Democracy 27*(1), 5–19.

Bischof, D., T. Allinger, M. Juratic, and K. V. S. Frederiksen (2023). (Mis-) Perceiving Support for Democracy: The Role of Social Norms for Democracies. *Working Paper*.

Braley, A., G. S. Lenz, D. Adjodah, H. Rahnama, and A. Pentland (2023). Why voters who value democracy participate in democratic backsliding. *Nature Human Behaviour*, 1–12. Publisher: Nature Publishing Group.

Braman, E. (2021). Thinking about Government Authority: Constitutional Rules and Political Context in Citizens' Assessments of Judicial, Legislative, and Executive Action. *American Journal of Political Science 65*(2), 389–404. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12582.

Chiopris, C., M. Nalepa, and G. Vanberg (2021). A wolf in sheep's clothing: Citizen uncertainty and democratic backsliding.

Cinar, I. and M. Nalepa (2022). Mass or Elite Polarization as the Driver of Authoritarian Backsliding? Evidence from 14 Polish Surveys (2005–2021). *Journal of Political Institutions and Political Economy 3*(3–4), 433–448.

Clayton, K., N. T. Davis, B. Nyhan, E. Porter, T. J. Ryan, and T. J. Wood (2021). Elite rhetoric can undermine democratic norms. *Proceedings of the National Academy of Sciences 118*(23), e2024125118.

Das, S. (2023). Democratic backsliding in the world's largest democracy. *Available at SSRN 4512936*.

De Bruin, E. (2018). Preventing coups d'état: How counterbalancing works. *Journal of Conflict Resolution 62*(7), 1433–1458.

Druckman, J. N., S. Kang, J. Chu, M. N. Stagnaro, J. G. Voelkel, J. S. Mernyk, S. L. Pink, C. Redekopp, D. G. Rand, and R. Willer (2023). Correcting misperceptions of out-partisans decreases american legislators' support for undemocratic practices. *Proceedings of the National Academy of Sciences 120*(23), e2301836120.

Frederiksen, K. V. S. (2022a). Does Competence Make Citizens Tolerate Undemocratic Behavior? *American Political Science Review 116*(3), 1147–1153.

Frederiksen, K. V. S. (2022b). More Parties, More Punishment of Undemocratic Candidates? Experiments on England's Ambiguous Party System. preprint, Politics and International Relations.

Gamboa, L. (2017). Opposition at the margins: Strategies against the erosion of democracy in colombia and venezuela. *Comparative Politics 49*(4), 457–477.

Gandhi, J. (2019). The institutional roots of democratic backsliding. *The Journal of Politics 81*(1), e11–e16.

Gandhi, J. and E. Ong (2019). Committed or Conditional Democrats? Opposition Dynamics in Electoral Autocracies. *American Journal of Political Science 63*(4), 948–963. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12441.

Geddes, B. (1999). What do we know about democratization after twenty years? *Annual review of political science 2*(1), 115–144.

Gibler, D. M. and K. A. Randazzo (2011). Testing the Effects of Independent Judiciaries on the Likelihood of Democratic Backsliding. *American Journal of Political Science 55*(3), 696–709.

Gidengil, E., D. Stolle, and O. BERGERON-BOUTIN (2022). The partisan nature of support for democratic backsliding: A comparative perspective. *European Journal of Political Research 61*(4), 901–929.

Gratton, G. and B. E. Lee (2023). Liberty, Security, and Accountability: The Rise and Fall of Illiberal Democracies. *The Review of Economic Studies*.

Green, D. P., P. M. Aronow, D. E. Bergan, P. Greene, C. Paris, and B. I. Weinberger (2011). Does Knowledge of Constitutional Principles Increase Support for Civil Liberties? Results from a Randomized Field Experiment. *The Journal of Politics 73*(2), 463–476.

Grillo, E. and C. Prato (2023). Reference points and democratic backsliding. *American Journal of Political Science 67*(1), 71–88.

Grossman, G., D. Kronick, M. Levendusky, and M. Meredith (2022). The Majoritarian Threat to Liberal Democracy. *Journal of Experimental Political Science 9*(1), 36–45.

Grumbach, J. M. (2022). Laboratories of Democratic Backsliding. *American Political Science Review*, 1–18.

Haggard, S. and R. Kaufman (2021). *Backsliding: Democratic regress in the contemporary world*. Cambridge University Press.

Helmke, G., M. Kroeger, and J. Paine (2022). Democracy by deterrence: Norms, constitutions, and electoral tilting. *American Journal of Political Science 66*(2), 434–450.

Hollyer, J. R., M. Klašnja, and R. Titiunik (2023). Charismatic leaders and democratic backsliding.

Hollyer, J. R., B. P. Rosendorff, J. R. Vreeland, and J. R. Hollyer (2011). Democracy and Transparency. *The Journal of Politics 73*(4), 1191–1205. Publisher: The University of Chicago Press.

Howell, W. G., K. A. Shepsle, and S. Wolton (2023). Executive absolutism: The dynamics of authority acquisition in a system of separated powers. *Quarterly Journal of Political Science 18*(2), 243–275.

Howell, W. G. and S. Wolton (2018). The Politician's Province. *Quarterly Journal of Political Science 13*(2), 119–146.

Huntington, S. P. (1993). *The third wave: Democratization in the late twentieth century*, Volume 4. University of Oklahoma press.

Kingzette, J., J. N. Druckman, S. Klar, Y. Krupnikov, M. Levendusky, and J. B. Ryan (2021). How affective polarization undermines support for democratic norms. *Public Opinion Quarterly 85*(2), 663–677.

Krishnarajan, S. (2023). Rationalizing Democracy: The Perceptual Bias and (Un)Democratic Behavior. *American Political Science Review 117*(2), 474–496.

Kuran, T. (1991). Now out of never: The element of surprise in the east european revolution of 1989. *World politics 44*(1), 7–48.

Levitsky, S. and D. Ziblatt (2018). *How Democracies Die*. Crown. Google-Books-ID: VZKADwAAQBAJ.

Linz, J. J. and A. C. Stepan (1978). *The breakdown of democratic regimes, Latin America*. (Johns Hopkins University Press).

Little, A. and A. Meng (2023). Measuring democratic backsliding. *Available at SSRN 4327307*.

Luo, Z. and A. Przeworski (2023). Democracy and its vulnerabilities: Dynamics of democratic backsliding. *Quarterly Journal of Political Science 18*(1), 105–130.

Mazepus, H. and D. Toshkov (2021). Standing up for Democracy? Explaining Citizens' Support for Democratic Checks and Balances. *Comparative Political Studies* (8).

Meng, A. (2020). *Constraining Dictatorship*. Cambridge University Press Cambridge.

Miller, M. K. (2021). A republic, if you can keep it: Breakdown and erosion in modern democracies. *The Journal of Politics 83*(1), 198–213.

Orhan, Y. E. (2022). The relationship between affective polarization and democratic backsliding: comparative evidence. *Democratization 29*(4), 714–735.

O'donnell, G., P. C. Schmitter, and L. Whitehead (2013). *Transitions from authoritarian rule: Tentative conclusions about uncertain democracies*. JHU Press.

Pospieszna, P. and A. Vetulanicegiel (2021). Polish interest groups facing democratic backsliding. *Interest Groups & Advocacy 10*(2), 158–180.

Przeworski, A. (1991). *Democracy and the market: Political and economic reforms in Eastern Europe and Latin America*. Cambridge university press.

Przeworski, A. (2000). *Democracy and development: Political institutions and well-being in the world, 1950-1990*. Number 3. Cambridge University Press.

Przeworski, A. (2019). *Crises of democracy*. Cambridge University Press.

Pérez-Liñán, A., N. Schmidt, and D. Vairo (2019). Presidential hegemony and democratic backsliding in Latin America, 1925–2016. *Democratization 26*(4), 606–625. Publisher: Routledge _eprint: https://doi.org/10.1080/13510347.2019.1566321.

Rovny, J. (2023). Antidote to backsliding: Ethnic politics and democratic resilience. *American Political Science Review*, 1–19.

Saikkonen, I. A.-L. and H. S. Christensen (2023). Guardians of democracy or passive bystanders? a conjoint experiment on elite transgressions of democratic norms. *Political Research Quarterly 76*(1), 127–142.

Simonovits, G., J. McCoy, and L. Littvay (2022). Democratic Hypocrisy and Out-Group Threat: Explaining Citizen Support for Democratic Erosion. *The Journal of Politics 84*(3), 1806–1811.

Svolik, M. (2021). Voting Against Autocracy.

Svolik, M. W. (2012). *The politics of authoritarian rule*. Cambridge University Press.

Svolik, M. W. (2020). When Polarization Trumps Civic Virtue: Partisan Conflict and the Subversion of Democracy by Incumbents. *Quarterly Journal of Political Science 15*(1), 3–31.

Svolik, M. W., E. Avramovska, J. Lutz, and F. Milaèiæ (2023). In europe, democracy erodes from the right. *Journal of Democracy 34*(1), 5–20.

Thompson, A. (2021). How Racial Threat Motivates Partisan Differences in Anti-Democratic Attitudes.

Urbinati, N. (2019). Political theory of populism. *Annual review of political science 22*, 111–127.

Vachudova, M. A. (2021). Populism, democracy, and party system change in europe. *Annual Review of Political Science 24*, 471–498.

Van Noort, S. (2022). How Strongly Do American Voters React to Anti-Democratic Behavior by Politicians? Natural Experimental Evidence from the January 6 Insurrection. preprint, Politics and International Relations.

Waldner, D. and E. Lust (2018). Unwelcome change: Coming to terms with democratic backsliding. *Annual Review of Political Science 21*, 93–113.

Wunsch, N., M. Jacob, and L. Derksen (2023). The Demand Side of Democratic Backsliding: How Divergent Understandings of Democracy Shape Political Choice. *Working Paper*.

Wuttke, A. and F. Foos (2022). Making the Case for Democracy. *Working Paper*.

Wuttke, A., C. Schimpf, and H. Schoen (2023). Populist Citizens in four European Countries: Widespread Dissatisfaction goes with Contradictory but Pro-democratic Regime Preferences. *Swiss Political Science Review*.

Wuttke, A., F. Sichart, and F. Foos (2023). Null Effects of Pro-Democracy Speeches by U.S. Republicans in the Aftermath of January 6th. *JEPS*.

Ziblatt, D. (2017). *Conservative Political Parties and the Birth of Modern Democracy in Europe*. Cambridge University Press. Google-Books-ID: leCBDgAAQBAJ.